Clinical Interpretable Deep Learning Model for Glaucoma Diagnosis

WangMin Liao, BeiJi Zou, RongChang Zhao, YuanQiong Chen, ZhiYou He, and MengJie Zhou

***Abstract*—Despite the potential to revolutionise disease diagnosis by performing data-driven classification, clinical interpretability of ConvNet remains challenging. In this paper, a novel clinical interpretable ConvNet architecture is proposed not only for accurate glaucoma diagnosis but also for the more transparent interpretation by highlighting the distinct regions recognised by the network. To the best of our knowledge, this is the first work of providing the interpretable diagnosis of glaucoma with the popular deep learning model. We propose a novel scheme for aggregating features from different scales to promote the performance of glaucoma diagnosis, which we refer to as M-LAP. Moreover, by modelling the correspondence from binary diagnosis information to the spatial pixels, the proposed scheme generates glaucoma activations, which bridge the gap between global semantical diagnosis and precise location. In contrast to previous works, it can discover the distinguish local regions in fundus images as evidence for clinical interpretable glaucoma diagnosis. Experimental results, performed on the challenging ORIGA datasets, show that our method on glaucoma diagnosis outperforms state-of-the-art methods with the highest AUC (0.88). Remarkably, the extensive results, optic disc segmentation (dice of 0.9) and local disease focus localization based on the evidence map, demonstrate the effectiveness of our methods on clinical interpretability.**

***Index Terms*—Glaucoma diagnosis, clinical interpretation, medical image processing.**

I. Introduction

GLAUCOMA is a major chronic eye disease that acts as the second leading cause of blindness worldwide, with around 80 million people by 2020 [1], [2]. Since glaucoma can cause irreversible vision loss, early diagnosis is critical to slow down the progress [3]. Clinically, the usual diagnosis includes intra-ocular pressure and visual field loss tests together with a manual assessment of the optic disc (OD) through ophthalmoscopy. However, it is difficult and time-consuming for manual detection due to its complex procedure. As shown in Fig. 1, manual measurement is always required to quantificationally assess the structural changes and progressive damage of optical nerve head (ONH) caused by glaucoma [4]. In clinical practice, widely-adopted quantitative measurements include cup-to-disc ratio (CDR), rim to disc area ratio, disc diameter, disc area and so on [5]. Besides, the notch on neuroretinal rim [6], the bleeding on optic disc [7] and defects on retinal nerve fibre layer [8] are employed as evidence to provide detail information for accurate assessment of ONH. Therefore, the clinical evidences of glaucoma are distributed on the OD.

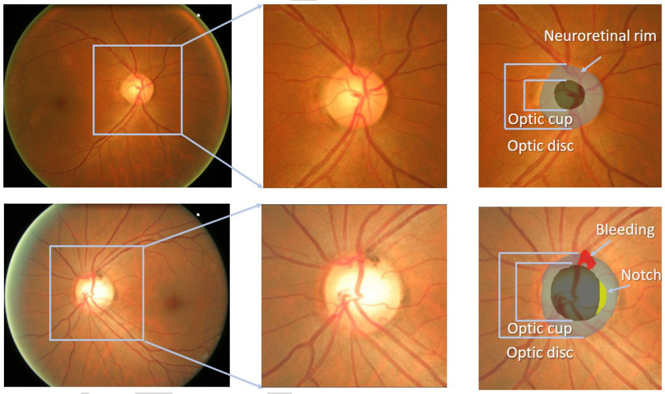


Fig. 1. Top: a normal image. Bottom: a glaucoma image. The glaucoma image has a higher cup-to-disc ratio (CDR). And some of them have bleeding spots and notch on the neuroretinal rim. They are the evidence for glaucoma diagnosis.

Nowadays, convolutional neural networks (CNN) based ONH assessment methods have been widely used for large-scale automated diagnosis of glaucoma [9]–[14]. With the rapid development of medical imaging [15], [16], these machine learning methods make rapid diagnosis possible, and it is significant for the screening in community health centres [17]–[18]. Although these methods make breakthroughs in automated glaucoma diagnosis, they still suffer from some weakness. The most criticised one is lack of clinical interpretation and explicit diagnosis evidence [19]–[21]. CNN-based methods can often provide diagnostic conclusions accurately. However, they cannot bring out the facts or reasons why the conclusions are made. To solve this problem, we provide a pathological condition for physicians and intuitive interpretation for patients of how the diagnosis made, as clinical evidence. In a computer-aided

Manuscript received February 19, 2019; revised July 1, 2019 and August 23, 2019; accepted October 15, 2019. This work was supported in part by the National Natural Science Foundation of China (61702558 and 61573380); in part by the Hunan Natural Science Foundation (2017JJ3411); in part by Key R&D projects in Hunan (2017WK2074); in part by the National Key R&D Program of China (2017YFC0840104); and in part by the Graduate Research and Innovation Project of Central South University (2019zzts587). *(Corresponding author: RongChang Zhao.)*

The authors are with the School of Computer Science and Engineering, Central South University, Changsha 410083, China, and also with the Hunan Engineering Research Center of Machine Vision and Intelligent Medicine, Changsha 410012, China (e-mail: 0909123117@csu.edu.cn; bjzou@csu.edu.cn; byrons.zhao@gmail.com; yqchen@vip.163.com; freehe@csu.edu.cn; 13017384616@163.com).

Digital Object Identifier 10.1109/JBHI.2019.2949075

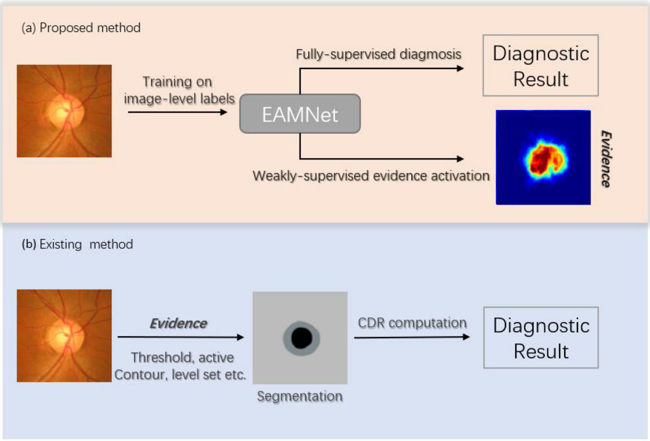

(a) Proposed method
Training on image-level labels
EAMNet
Fully-supervised diagmosis
Diagnostic Result
Weakly-supervised evidence activation
Evidence
(b) Existing method
Evidence
Threshold, active Contour, level set etc.
Segmentation
CDR computation
Diagnostic Result

Fig. 2. (a) The proposed method not only obtains automated diagnostic conclusions but also provides the clinical evidence for the accurate diagnosis. (b) Traditional segmentation-based methods first measure CDR from the segmented result which requires strong prior information and user interaction, then diagnosis glaucoma based on the segmentation results.

approach, the clinical evidence of glaucoma is often shown as changes in intensity or structure in local regions. Unfortunately, modern CNN has difficulties in dealing with the problem of evidence identification. It is because we use CNN as a black box. The clinical evidence is hidden in the black box. It is the most challenging task to bridge the gap between the evidence of a model and the understanding of the ophthalmologist. Due to the pyramid structure of a CNN, the flow of information and the region of interest are imperceptible. Since we can not open an old box, we can make an openable box.

Designing a system, which provides reliable evidence for accurate diagnosis of glaucoma is an excellent challenging task in clinical practice [22]. For the ophthalmologist, the clear and easy to be understood evidence for the glaucoma diagnosis is the localization of lesions. Existing methods usually treat evidence extraction and glaucoma diagnosis as two separate tasks that are solved with two independent systems. On the one hand, many methods [9], [23]–[26] have been proposed to find the evidence area by localizing and segmenting the anatomies with supervised technique. However, those segmented areas are not always sensitive to the accurate diagnosis as pathological conditions. On the other hand, glaucoma diagnosis is formulated as a classification problem in machine learning to be solved end-to-end [9]–[13]. The classification model is a black box, and neither clinician nor patients can be told why it is, but only what it is. Multi-task learning [10] have been used to find the segmented area and diagnosis glaucoma simultaneously. However, multi-task learning usually needs large-scale pixel-level annotation which is expensive to obtain. Weakly-supervised learning has the ability to find local special regions only with classification labels [27]. Fig. 2 demonstrates how the proposed method differs from segmentation-based methods of getting the evidence.

To make it clear and easy to be understood, the evidence should be highly correlated to diagnosis. In fact, for the ophthalmologist, the clinical evidence for the glaucoma diagnosis is the segmentation of optic disc and cup along with localization of lesions. If the region of interest matches the clinical evidence area, optic disc and lesions, the model is interpretable. In this paper, we propose a novel clinical interpretable ConvNet architecture (EAMNet) not only to achieve accurate glaucoma diagnosis but also to provide a more transparent interpretation by highlighting the distinct regions recognized by the network. Therefore, our EAMNet enables deep model interpretable benefitting from three facts: 1) the model imitates the diagnosis process of clinical physicians who discover the evidence to support the diagnosis. The proposed EAMNet not only gives the diagnosis results, but also provides a visual region of interest (ROI) to corroborate the reliability of the diagnosis decision. 2) the proposed EAMNet employs three distinguished components to accurately discover local regions with particular appearance and features to support the glaucoma diagnosis. Specifically, a well-designed CNN has constructed to abstract hierarchical information for semantic features extraction and automated glaucoma diagnosis. A novel method, Multi-Layers Average Pooling (M-LAP), is proposed to build an information passageway to bridge the gap between semantic information and localization information at multiple scales. 3) the results produced by our EAMNet are interpretable for glaucoma diagnosis due to it can discover ophthalmic lesions and key anatomical regions (OD) automatically without any pixel-level annotation, as shown in Section III. The contribution of our work is as follows:

1) For the first time, a clinical interpretable deep learning model is proposed to not only achieve accurate automated glaucoma diagnosis but also provide a more transparent interpretation by highlighting the distinct regions to support the diagnosis.
2) A novel method, Multi-Layers Average Pooling (M-LAP), is proposed to integrate features of different levels for accurate glaucoma diagnosis, meanwhile building an information passageway to bridge the gap between semantic information and localization information at multiple scales and collaborating with Evidence Activation Mapping this method both output fully-supervised diagnosis and weakly-supervised evidence localization.
3) We achieve clinical interpretable diagnosis result of high accuracy. Our method on glaucoma diagnosis achieves state-of-the-art accuracy with the Area Under Curve (AUC) of 0.88, and it provides the evidence activation maps which give the clinical basis of glaucoma, which is meaningful for the clinical application of CNN.

II. Methodology

The proposed framework (EAMNet), as shown in Fig. 3, mainly consists of three main parts: CNN backbone network for hierarchical feature extraction and aggregation, Multi-Layers Average Pooling (M-LAP) to bridge the gap between semantic information and localization information at multiple scales and Evidence Activation Mapping for evidence identification and discovery. We adopt a classification network with ResBlock and multiple convolutional layers as a backbone network, which obtains excellent representation via aggregation of complex

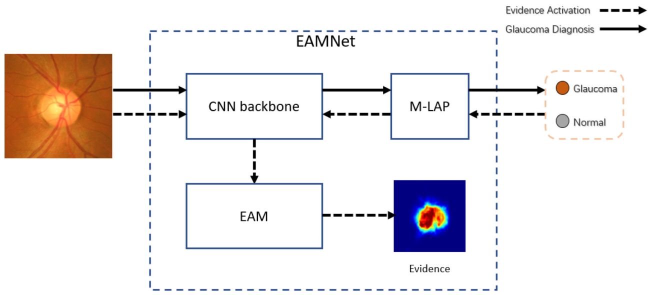
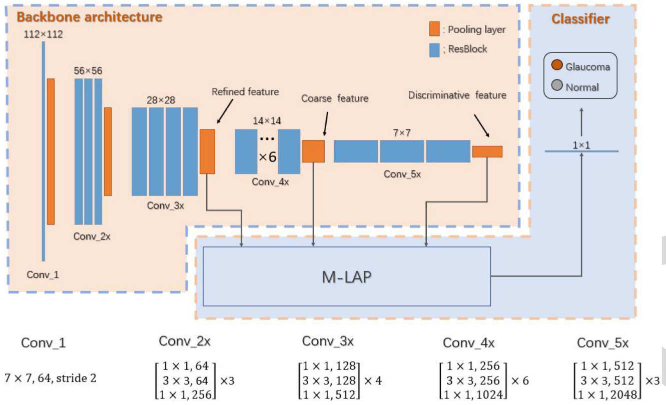


Fig. 3. The overview of the proposed EAMNet, containing three main parts: CNN backbone network for hierarchical feature extraction and aggregation, Multi-Layers Average Pooling (M-LAP) to bridge the gap between semantic information and localization information at multiple scales and Evidence Activation Mapping for evidence identification and discovery.



Conv_1	Conv_2x	Conv_3x	Conv_4x	Conv_5x
$7 \times 7, 64, \text{stride } 2$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$

Fig. 4. The overview of backbone architecture is shown in the yellow box. The input image is a 224 × 224 RGB image. There are five stages in the architecture. Each stage includes several ResBlocks and one pooling layers. The pooling layer is 2 × 2 max pooling. The architecture of ResBlocks in different stages are on the bottom. The output of each selected stage is connected to the next stage and also followed by a 1 × 1 convolution layer to decrease the parameter size. The chosen stage is the input of M-LAP.

hierarchical features. To bridge the gap between semantic information and localization information, we perform a novel block M-LAP on the convolutional hierarchical feature maps. The method produces the diagnosis conclusion. Meanwhile, it provides evidence activation map. Then EAM projects back the binary diagnosis conclusion on to the convolutional feature maps and activates the local pixels which contribute to glaucoma diagnosis. Therefore, EAMNet can discover and identify the particular local regions of the fundus image (notch on the neuroretinal rim, bleeding on optic disc and defects on the optic disc, etc.).

A. Backbone Architecture

The backbone of EAMNet is a feature expressive representations network with multiple convolutional layers and pooling layers. As shown in Fig. 4. We use ResBlock [29] as the basic module of our network. These ResBlocks are connected to different ResBlocks or pooling layers. We select three pooling layers according to the different levels of feature layers. We resize the output of these pooling layers and concatenate them. The identity shortcut connection it introduces provides a fairly good representation of fundus images, which largely enhances the ability to extract evidence and the ability to diagnose glaucoma. Considering the spatial layout of fundus images are almost the same and avoiding model redundancy, we configure a low number of filters. We also largely employ dropout and batch normalisation layers to alleviate overfitting. Just before sent into the network, a fundus image is resized to 224 × 224. As can be seen in our experiments, our CNN architecture is beneficial for fundus images representation.

Noting that there are five stages in the architecture. Each stage includes several ResBlocks and one pooling layers. The first stage, Conv_1, include a 7 × 7 convolution layer. Others are ResBlocks with three convolution layers and shortcut connection. As shown in Fig. 4, the architecture of different stages are slightly different in the size of output and the repeat times of ResBlocks. Experimentally, we select three stages as the input of M-LAP, which are Conv_3x, Conv_4x and Conv_5x stages. Their pooling layers are followed by a 1 × 1 convolution layer to decrease the parameter size. And they are resized to the same size in M-LAP.

Each ResBlock is a combination of convolution layers. The architecture explicitly enables each layers fit a residual mapping instead of letting each few stacked layers directly fit a desired underlying mapping. Denoting the underlying mapping as $H(x)$, the stacked nonlinear layers fit another mapping of $F(x) = H(x) - x$. This is the formulation of a shortcut connection. The origin mapping $H(x)$ is $F(x) + x$. It shows that identity shortcut connections add neither extra parameters nor computational complexity [29].

B. Multi-Layers Average Pooling

We introduce the Multi-Layers Average Pooling (M-LAP) to aggregate multi-scale global features for glaucoma diagnosis effectively. Meanwhile, the M-LAP provides an information passageway to bridge the gap between semantic information and localization information at multiple scales. As shown in Fig. 5, M-LAP consists of multi-scale feature aggregation and channel-wise global pooling. With the multi-scale feature maps, the mission is constructing a classifier to achieve accurate classification between glaucoma and normal image by aggregating feature maps. Given the fact that the lesions of glaucoma are of different layout and size. In our implements, three-level feature maps, refined, coarse and discriminative features, are aggregated to obtain expressive representations of fundus images. To aggregate feature in different scales, we first resize all feature maps to the same size as the output of the feature extractor. Then all the resized feature maps are concatenated into a multi-channel feature map followed by the 1 × 1 convolution to interactively aggregate features among different channels and generate fix-channel feature maps.

Different from the traditional classifier with fully-connection layer, M-LAP uses the global spatial pooling to abstract the semantics for accurate classification. Global spatial pooling (GSP) averages the feature map into represented single value instead of every pixels. As shown in Fig. 6 , the GSP [30] layer is simple

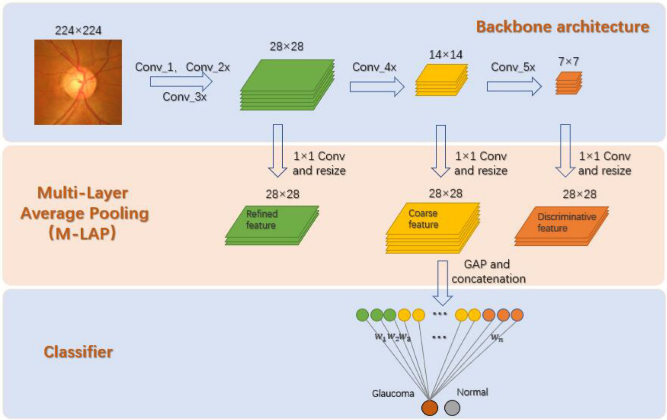


Fig. 5. The three modules are connected in this way. The selected stages of backbone architecture are connected to the M-LAP. Three branches with 1 × 1 convolution and resizing are used to modify the size of feature maps to concatenate them. After this process, three feature maps are generated, which represent refined features, coarse features and discriminative features respectively. Before classification, global average pooling is used to generate one-dimensional vectors.

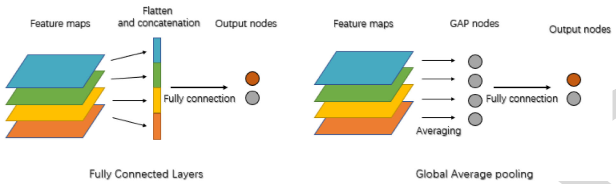


Fig. 6. Fully connected layers flatten the feature maps. The network contains lots of redundant information, making it impossible to project information from arrays to feature maps. Global average pooling global average each feature maps as the representation of them.

in structure and needs fewer parameters to train. For a given feature map, let $f_{ki}(x, y)$ represent the activation of channel k in layer i at (x, y). Then, for channel k in activation layer i, the result of global spatial pooling, F_{ki} is $\Sigma_{x,y} f_{ki}(x\,y)$. Thus, the output of the softmax layer of a given class c, $S_c = \Sigma_{k,i} w^c_{ki} F_{ki}$ where w^c_{ki} is the weight corresponding to class c for channel k in activation layer i.

By plugging $F_{ki} = \Sigma_{x,y} f_{ki}(x\,y)$ into the class score, S_c, we obtain:

$$S_c = \Sigma_{x,y} \Sigma_{k,i} w^c_{ki} f_{ki}\,(x\,y)\,. \tag{1}$$

It is easy to find out that the number of $\Sigma_{x,y} f_{ki}(x\,y)$ is the same as that of w^c_{ki} and the number of concatenated feature maps, which makes it possible to project weights back onto feature maps. Comparing to fully connection layer, the parameters are reduced by $1/xy$.

C. Evidence Activation Mapping

It is known that the shallower layers represent low-semantics features while the deeper layers represent discriminative features in a classification-oriented CNN [31]. Meanwhile, the shallower layers provide rich spatial information with high-resolution feature maps. Due to the structure of CNN, which is known as pyramid structure, discriminative feature maps are shrunk to an unacceptable small size at the deeper layers. This bottle obstructs the generation of feature maps with accurate spatial information and high semantics. In EAMNet, we describe a novel approach to generate refined evidence activation maps (EAM) with accurate spatial evidence information for glaucoma diagnosis.

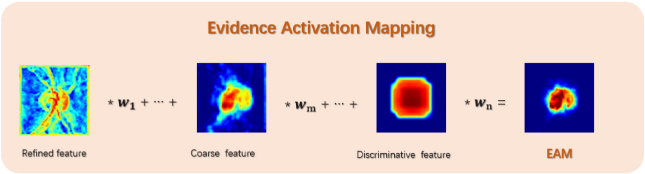


Fig. 7. Optic Disc Activation Mapping: the weights are mapped back to the previous convolutional layer to generate the Evidence Activation Maps (EAMs) as the attention score for glaucoma classification. There are n feature maps in all of the feature maps. Correspondingly there are n weights learned from the previous process. Weighted summation of weights and feature maps are used to generate EAM. The EAM highlights the glaucoma-specific discriminative regions.

Evidence activation mapping is a channel-wise attention-based approach for evidence identification and implemented by a projection from binary classification to spatial evidence maps. As shown in Fig. 7 the feature maps at different scales are aggregated into a single map by a weighted sum function, and the weighted sum function acts as an attention gate which gives the biggest weight to the feature map that contributes to glaucoma classification while giving small weight to the other one. Here, the weights are regarded as the attention scores for glaucoma classification and optimised in the classification stage. With the weighted sum function, EAMNet back-projects the attention scores from glaucoma classification to the different feature maps. In this implementation, we compute a weighted sum of the feature maps from three chosen convolutional layers to obtain our EAM. Let $g_{ki}(x\,y)$ represents the result of normalization of the kth kernel in the ith activation layer, where $(x\,y)$ is the coordinate of a pixel. In our method, there are 3 activation layers, as shown in Fig. 5. They are refined layers, coarse layers and discriminative layers. Each feature map, $g_{ki}(x\,y)$, has the same size of 28 × 28. We define M_c as the evidence activation map where the optic disc region share the same location with the significant evidence for glaucoma diagnosis.

$$M_c\,(x\,y) = \sum_i \sum_k w^c_{ki} g_{ki}\,(x\,y)\,. \tag{2}$$

where $f_{ki}(x\,y)$ is the feature map of the kth kernel in the ith activation map and w^c_{ki} is the weight learned by the classifier as is shown in Fig. 5. The kernel k and activation layer i are corresponding to the feature maps $g_{ki}(x, y)$.

The further experiments, which will be discussed in Section III-B, indicate that the multi-scale feature maps concatenation algorithm performs better than single-scale feature maps for evidence identification. It is because multi-scale feature maps provide more detailed spatial information of evidence at multiple scales. Our EAMNet makes the evidence map sharp and clear by using refined features while enhancing the semantics of evidence map by using coarse and discriminative features. The lesions in the optic disc are accurately discovered due to the three kinds of features.

III. Experiment

The effectiveness of the proposed EAMNet is validated on two aspects: the accuracy of glaucoma diagnosis and precision of evidence identification. We perform experiments on the challenging public datasets ORIGA [34]. The experimental results verify the proposed EAMNet achieves state-of-the-art diagnosis accuracy (0.88) and does an excellent performance on evidence identification.

In our experiments, the localization of lesions and segmentation of the optic disc are employed as an instance of evidence identification for our clinical interpretable EAMNet. The pathogenesis of glaucoma, structural changes of optical nerve head, are often observed on the optic disc [1]. It is believed that when judging a fundus image, whether it is glaucoma, doctors focus mostly on the optic disc and the lesions on it. Thus, when a CNN model provides diagnosis result, meanwhile giving the evidence map where the optic disc is, we are convinced this model is clinically interpretable. In this implementation, we make use of superpixel to soften the gradient of local features and employ ellipse fitting to obtain the segmentation of optic disc. To the best of our knowledge, no previous work sets a criterion to measure the interpretability of the model.

A. Criteria

In this paper, we utilise the area under the curve (AUC) of the receiver operation characteristic curve (ROC) to evaluate the performance of glaucoma diagnosis. The ROC is plotted as a curve which shows the tradeoff between sensitivity (TPR) and specificity (TNR), defined as:

$$\text{TPR} = \frac{TP}{TP+FN}, \quad \text{TNR} = \frac{TN}{FP+TN}. \tag{3}$$

where TP and TN are the numbers of true positives and true negatives, and FP along with FN are the number of false positives and false negatives, respectively.

We utilize the overlapping error E and balance accuracy A as the evaluation metrics for optic disc segmentation.

$$E = 1 - \frac{\text{Area}(S \cap G)}{\text{Area}(S \cup G)}, \quad A = \frac{1}{2}(TPR + TNR) \tag{4}$$

with

$$\text{TPR} = \frac{TP}{TP+FN}, \quad \text{TNR} = \frac{TN}{FP+TN} \tag{5}$$

where S and G denote the segmented mask and the manual ground truth, respectively.

B. Dataset

The origa dataset is used in the experiments to validate glaucoma diagnosis, disc segmentation and lesion localization. The ORIGA datasets are comprised of 168 glaucoma and 482 normal images from studies of a Malay population with ground truth cup and disc labels along with clinical glaucoma diagnoses. It was conducted over three years from 2004 to 2007 by the Singapore Eye Research Institute and funded by the National Medical Research Council. Singapore Malay Eye Study (SiMES) examined 3,280 Malay adults aged 40 to 80, from which, 149 are glaucoma patients. Retinal fundus images for both eyes were taken for each subject in the study [34]. The 650 images with manual labelled optic disc mask are divided into 325 training images (including 73 glaucoma cases) and 325 testing images (including 95 glaucomas).

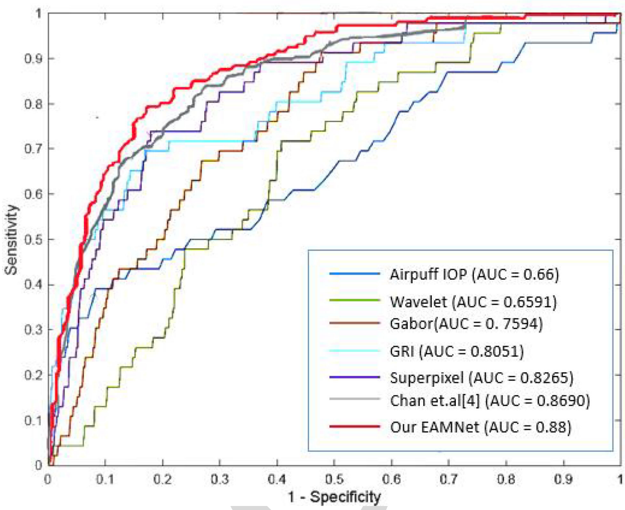


Fig. 8. ROC curve of our method and other methods. Our method, the red one, performs better than others.

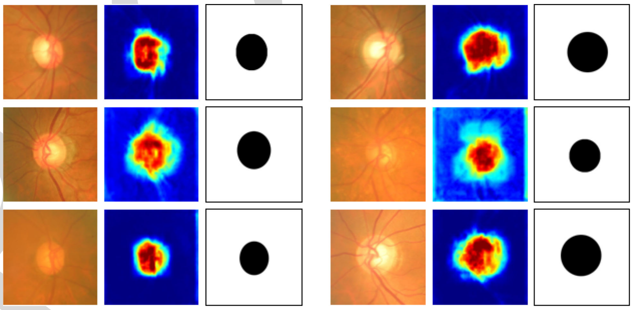


Fig. 9. The activated map represents optic disc and cup areas simultaneously with different activation amplitude. The first column is the raw fundus image, and the second and third columns are the activation map and segmented optic disc mask by our EAMNet.

1) Ablation Study: As shown in Figs. 8 and 9, the ablation study demonstrates that our method can not only obtain accurate glaucoma diagnosis but also provides the more transparent interpretation by highlighting the distinct regions recognised by the network. In Fig. 9, the ROC curve (in red) indicates that although the detection of glaucoma based on colour fundus image is a challenging task, our EAMNet obtains high sensitivity and low specificity. Thanks to the accurate evidence and multi-scale feature aggregation, EAMNet obtains a state-of-the-art AUC value with 0.88. It is much higher than the traditional image processing methods like Airpuff, Wavelet, Gabor, and GRI. It also performs better than the superpixel and CNN method (Chan *et al.* [4]). The further analytical result will be shown in

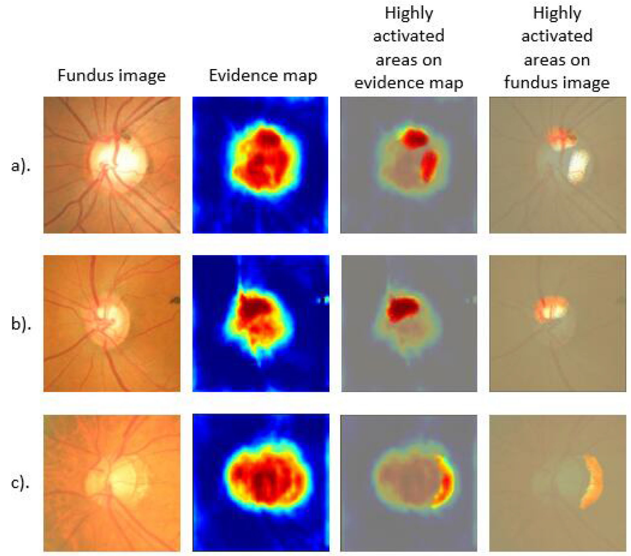

Fundus image
Evidence map
Highly activated areas on evidence map
Highly activated areas on fundus image
a).
b).
c).

Fig. 10. (a) Shows that notch and bleeding spots are highlighted on the final map. (b) Shows that the structural variation of blood vessels is also highlighted. (c) Indicates that PPA is taken into consideration of our method.

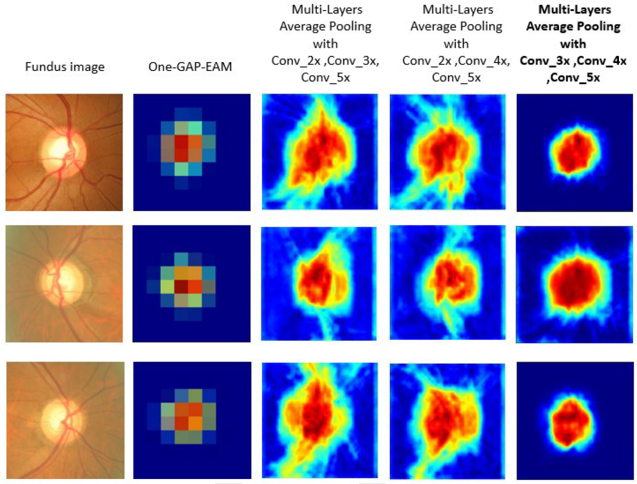

Fundus image
One-GAP-EAM
Multi-Layers Average Pooling with Conv_2x ,Conv_3x, Conv_5x
Multi-Layers Average Pooling with Conv_2x ,Conv_4x, Conv_5x
Multi-Layers Average Pooling with Conv_3x ,Conv_4x ,Conv_5x

Fig. 11. (1) Comparison of One-GAP-EAM and Multi-Layers Average Pooling. The results of One-GAP-EAM and Multi-Layers Average Pooling are shown in the middle and third columns, respectively. It is a 7 × 7 map which only shows the approximate position of the optic disc. The resolution is not enough for segmentation. The result of our proposed method is shown in the right row, which uses the feature maps of 28 × 28, 14 × 14 and 7 × 7. The final EAM is much finer. (2) We also test the result when EAM inserted in different layers. We find that when shallower layers are inserted, the result will be affected by the identity information, like vessels and texture of retina.

TABLE I
RELATIONSHIP OF CLASSIFICATION AND SEGMENTATION

Q2

iteration	AUC	A_{disc}
3	0.58	0.05
80	0.69	0.23
200	0.73	0.72
1600	**0.88**	**0.90**
50000(overfitted)	0.99(training set)	0.84

the comparison results part. Different from existing methods, EAMNet develops a novel technique named as multi-layer average pooling to extract discriminative features by aggregating multi-scale information strictly related to glaucoma diagnosis. This strategy improves above 1.1% compared with the existing direct classification methods.

Significantly, as shown in Fig. 9, EAMNet provides the precise activation area which contributes to glaucoma diagnosis. In our experiments, the activation maps are used to localize the lesions and segment optic disc to validate the effectiveness of our EAMNet. In Fig. 9, the second columns show that EAM activates the attention area as the pathogenesis area in fundus images for glaucoma diagnosis. The third columns in Fig. 9 are the segmented optic disc with our EAMNet. We can observe that the EAMNet can deal with the challenging optic disc segmentation task even though the image-level labels are used for training our model. It is worth noticing that the existing methods always achieve state-of-the-art results based on the supervised model with pixel-level labels.

It should be noted that the optic disc area is often segmented to measure the structural changes for accurate glaucoma diagnosis. Those distinguish regions show that our EAMNet is focused on the area of the optic disc and its lesions where the pathogenesis of glaucoma are highlighted for the diagnosis of clinicians. Also, we evaluate Multi-Layers Average Pooling and Single-Layer Average Pooling and find out that the ability of evidence activation is largely enhanced.

Noting that the distribution of EAM is uneven as shown in Fig. 10 . There are some extra activated areas beyond the optic disc. We observe these areas alone. These areas are the characteristics of glaucomatous related lesions. Such as bleeding, notch, PPA and structural variation of blood vessels. They are also crucial clinical evidence for glaucoma diagnosis. They are not always visible on the fundus images of glaucoma patients. We infer that our proposed method refers to not only the parameters of the optic disc and cup but also some rare features in the diagnosis of glaucoma. These features are also very important clinically, sometimes decisive. Therefore, we are convinced that the diagnostic basis of our method is the same as that of humans. It can be proven that our method is interpretable.

We compare the One-GAP-EAM model with Multi-Layers Average Pooling. As shown in Fig. 11, the result of One-GAP-EAM is not good enough to be used to segment optic disc. And, there is no other lesion area shown on the final One-GAP-EAM. Therefore, the EAM composed of feature maps with different resolutions can be better used to diagnose glaucoma and extract glaucoma lesions comprehensively.

In addition, experiments are conducted to demonstrate the clinical interpretation changes when the EAM module is inserted in different layers. To ensure the depth of the network, the last stage, Conv_5x, is always connected by the EAM module. As shown in Fig. 11, the outputs of the random structure are more likely to be affected by the irrelevant information, like vessels and texture of retina. It is because the models are likely to overfit.

TABLE II
CLASSIFICATION ON THE ORIGA VALIDATION SET

Method	AUC
EAMNet	**0.88**
Gabor [23]	0.66
Wavelet [24]	0.66
GRI [25]	0.81
Superpixel [26]	0.83
Chen et al. [10]	0.83
Zhao et al. [9]	0.86

TABLE III
OPTIC DISC SEGMENTATION ON THE ORIGA VALIDATION SET

Method	A_{disc}	E_{disc}
EAMNet	**0.90**	**0.29**
Superpixel [26]	0.96	0.26
U-Net [35]	0.96	0.12
M-Net + PT [36]	0.98	0.07

When the representation ability of a model is weak, the identity information, like vessels and texture of retina will dominante. Therefore, it can be proved that our structure can well represent the pathology of glaucoma rather than overfitting the data set.

We underfit the EAMNet step by step to explore the relationship of glaucoma diagnosis and optic disc segmentation in the unified framework. We observed that the result of glaucoma diagnosis is improved with the increasing of optic disc segmentation. We remove the batch normalisation layers in each ResBlocks and change the dropout rate to 0.2 to overfit the model. It can be found that as the overfitted accuracy raises the segmentation accuracy drops. It can be proven that although it looks like two independent tasks, the optic disc segmentation and glaucoma diagnosis in a unified framework are strongly related. We are convinced that the segmentation of optic disc is guided by the procedure of glaucoma diagnosis, while the accurate glaucoma diagnosis is also promoted by effective segmentation of optic disc as evidence map.

2) Comparison Results: In this section, we compare the results of proposed EAMNet with different types of CNN architectures and show that our EAMNet obtains the state-of-the-art performance on glaucoma diagnosis. Same as above, to quantify the evidence activation, we compare the results of optic disc segmentation which is generated by evidence activation maps with a generic and straightforward segmentation method. The matched methods are as follow. Gabor [23] and wavelet [24] method use manual features with Support Vector Machine (SVM) classifier to get the diagnostic result.GRI [25] is a probabilistic two-stage classification method to extract the Glaucoma Risk Index (GRI) that shows a good glaucoma detection performance. Superpixel [26] method proposes optic disc and optic cup segmentation using superpixel classification for glaucoma screening. Chen *et al.* [10] and Zhao *et al.* [9] propose two CNN method both of them have good accuracy. Meanwhile, U-Net [32] and M-Net + PT [36] are optic disc segmentation method also using CNN.

In the experiment, the manual labels are adopted as the ground truth. 10-fold cross-validation method is used in the experiment. We divided all samples into ten parts, each containing equal proportions of glaucoma and normal individuals. Each time nine samples were used as training samples, and the remaining one was used as a test sample. Finally, each result was averaged to obtain the final diagnosis result. As shown in Tables II and III, experimental results show that the proposed EAMNet achieves accurate glaucoma diagnosis (0.88 AUC) and optic disc segmentation (0.9 Adisc and 0.278 Edisc). Here, EAMNet obtains precise boundaries of the optic disc and accurate glaucoma diagnosis simultaneously since the accurate segmentation of optic disc originates from accurate glaucoma diagnosis. In addition, the accurate segmentation (even evidence identification) promotes and verify the accuracy of glaucoma diagnosis. Compared with state-of-the-art methods, our EAMNet achieves accurate glaucoma diagnosis, meanwhile obtains high performance on evidence activation.

As shown in Table III, the results show that EAMNet deals effectively with the challenging task of optic disc segmentation, even though the pixel-level is unavailable. Noting that our method is worse than other methods in the task of optic disc segmentation. It is because that we did not use any pixel-level labels, and there is much less supervision information in our task than fully-supervised method. We only use the fully-supervised method for comparison. And the comparison results are only for reference to prove that our semi-supervised method is as effective as other methods. Although using the image-level labels, EAMNet performs closely to fully-supervised OD segmentation methods. This phenomenon indicates that the main pathological area of glaucoma is located in the optic disc, which matches domain knowledge of glaucoma. And considering intuition clinical evidence of glaucoma, like CDR, closely related to the optic disc, it is interpretable when CNN activation map covers it.

IV. Conclusion and Future Work

In this paper, we propose a novel clinical interpretable ConvNet architecture named EAMNet not only for accurate glaucoma diagnosis but also for the more transparent interpretation by highlighting the distinct regions recognized by the network. The EAMNet solves the lack of interpretability of CNN-based glaucoma diagnosis CAD system. Beside diagnosing glaucoma with high precision, the proposed EAMNet also gives an interpretation for diagnosis. It presents the ability of weakly-supervised optic disc segmentation. And it activates the extract glaucoma lesions like bleeding, notch, PPA and structural variation of blood vessels. The proposed EAMNet employed the ResNet and M-LAP. It consists of 3 GAPs connecting to 3 layers of the different resolution increasing the resolution of EAM significantly. The result shows that this method makes classification performance primarily preserved. And an additional function of optic disc segmentation is attached. We have demonstrated that our system produces high accuracy diagnosis and optic disc segmentation results on ORIGA dataset.

Based on this work, limitations and open questions are drawn. High-resolution feature maps are hard to be represented by GAP. Besides, the optic cup is also important and related to glaucoma diagnosis. Further studies need to be carried out to design a more empirical model to deal with the clear cup segmentation by weakly-supervised evidence exploring.

References

[1] E. Dervisevic, S. Pavljasevic A. Dervisevic, and S. Kasumovic, “Challenges in early glaucoma detection,” *Med. Arch.*, vol. 70, pp. 203–207, 2016.

[2] H A. Quigley and A T. Broman, “The number of people with glaucoma worldwide in 2010 and 2020,” *Brit. J. Ophthalmology*, vol. 90, pp. 262–267, 2006.

[3] M. C. Leskea, A. Heijl, L. Hyman, B. Bengtsson, and E. Komaroff, “Factors for progression and glaucoma treatment: The early manifest glaucoma trial,” *Current Opinion Ophthalmology*, vol. 15, pp. 102–106, 2004.

[4] W. Chan *et al.*, “Analysis of retinal nerve fiber layer and optic nerve head in glaucoma with different reference plane offsets, using optical coherence tomography,” *Dig. World Core Med. J.*, 2006.

[5] M. A. Fernandez-Granero *et al.*, “Automatic CDR estimation for early glaucoma diagnosis,” *J. Healthcare Eng.*, vol. 2017, pp. 1–14, 2017. Q3

[6] M. H. Tan *et al.*, “Automatic notch detection in retinal images,” in *Proc. IEEE Int. Symp. Biomed. Imag.* 2013. Q4

[7] S. Vernon, “How to screen for glaucoma,” *Practitioner*, vol. 239, pp. 257–260, 1995.

[8] J. B. Jonas *et al.*, “Ranking of optic disc variables for detection of glaucomatous optic nerve damage,” *Investigative Ophthalmology Vis. Sci.*, vol. 41, pp. 1764–1773, 2000.

[9] R. Zhao *et al.*, “Automatic detection of glaucoma based on aggregated multi-channel features,” *J. Comput.-Aided Des. Comput. Graph.*, vol. 29, pp. 998–1006, 2017.

[10] B. Zou *et al.*, “Classified optic disc localization algorithm based on verification model,” *Comput. Graph.*, vol. 70, pp. 281–287, 2018.

[11] X. Chen *et al.*, “Glaucoma detection based on deep convolutional neural network,” *Eng. Medicine Biol. Sci.*, 2015.

[12] P. C. D. Junior, A. Sarmento, and A. Sarmento, “A HW/SW embedded system for accelerating diagnosis of glaucoma from eye fundus images,” in *Proc. Int. Symp. Rapid Syst. Prototyping: Shortening Path Specification Prototype*, 2016, pp. 12–18. Q5

[13] T. J. Jun *et al.*, “2sRanking-CNN: A 2-stage ranking-CNN for diagnosis of glaucoma from fundus images using CAM-extracted ROI as intermediate input,” 2018, *arXiv:1805.05727*. [Online] Available: https://arxiv.org/abs/1805.05727 Q6

[14] S. Yousefi *et al.*, “Glaucoma progression detection using structural retinal nerve fiber layer measurements and functional visual field points,” *IEEE Trans. Biomed. Eng.*, vol. 61, no. 4, pp. 1143–1154, Apr. 2014. Q7

[15] Z. Gao *et al.*, “Motion tracking of the carotid artery wall from ultrasound image sequences: A nonlinear state-space approach,” *IEEE Trans. Med. Imag.*, vol. 37, no. 1, pp. 273–283, Jan. 2018.

[16] Z. Gao, H. Xiong, and X. Liu, “Robust estimation of carotid artery wall motion using the elasticity-based state-space approach,” *Med. Image Anal.*, vol. 37, pp. 1–21, 2017.

[17] K. Blekas, D. I. Fotiadis, and A. Likas, “Greedy mixture learning for multiple motif discovery in biological sequences,” *Bioinformatics*, vol. 19, pp. 607–617, 2003.

[18] K. Kourou, C. Papaloukas, and D. I. Fotiadis, “Modeling biological data through dynamic bayesian networks for oral squamous cell carcinoma classification,” in *Proc. World Congr. Med. Phys. Biomed. Eng.*, 2018, pp. 375–379.

[19] Q. Zhang *et al.*, “Interpreting CNN knowledge via an explanatory graph,” in *Proc. Thirty-Second Assoc. Advancement Artif. Intell. Conf. Artif. Intell.*, 2017.

[20] C. M. Olthoff *et al.*, “Noncompliance with ocular hypotensive treatment in patients with glaucoma or ocular hypertension: An evidence-based review,” *Ophthalmology*, vol. 112, pp. 953–961, 2005.

[21] H. I. Suk and D. Shen, “Alzheimers Disease Neuroimaging Initiative. Deep learning in diagnosis of brain disorders,” *Deep Learning in Diagnosis of Brain Disorders*, The Netherlands: Springer, 2015, pp. 203–213. Q8

[22] R. Zhao *et al.*, “Weakly-supervised simultaneous evidence identification and segmentation for automated glaucoma diagnosis,” in *Proc. Thirty-Third Assoc. Advancement Artif. Intell. Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 01, pp. 809–816.

[23] U. R. Acharya *et al.*, “Decision support system for the glaucoma using Gabor transformation,” *Biomed. Signal Process. Control*, vol. 15, pp. 18–26, 2015.

[24] S. Dua, U. R. Acharya, P. Chowriappa, and S. V. Sree, “Wavelet-based energy features for glaucomatous image classification,” *IEEE Trans. Inf. Technol. Biomedicine*, vol. 16, no. 1, pp. 80–87, Jan. 2012.

[25] R. Bock, J. R. Meier, and L. G. Nyl, “Glaucoma risk index: Automated glaucoma detection from color fundus images,” *Med. Image Anal.*, vol. 14, pp. 471–481, 2010.

[26] J. Cheng *et al.*, “Superpixel classification based optic disc and optic cup segmentation for glaucoma screening,” *IEEE Trans. Med. Imag.*, vol. 32, no. 6, pp. 1019–1032, Jun. 2013.

[27] Z. Wang *et al.*, “Zoom-in-Net: Deep mining lesions for diabetic retinopathy detection,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2017, pp. 267–275.

[28] S. Barratt, “InterpNET: Neural introspection for interpretable deep learning,” *Preprint*, 2017, *arXiv:1710.09511*.

[29] K. He *et al.*, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 770–778.

[30] B. Zhou *et al.*, “Learning deep features for discriminative localization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[31] J. Ngiam *et al.*, “Sparse filtering,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1125–1133.

[32] M. Oquab *et al.*, “Is object localization for free? - Weakly-supervised learning with convolutional neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 685–694.

[33] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[34] Z. Zhang *et al.*, “ORIGA-light : An online retinal fundus image database for glaucoma analysis and research,” in *Proc. IEEE Conf. Eng. Med. Biol. Soc.*, 2009, vol. 2010, pp. 3065–3068.

[35] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[36] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, “Joint optic disc and cup segmentation based on multi-label deep network and polar transformation,” *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018.